Propagating Fine-Grained Topic Labels in News Snippets

Luís Sarmento*†, Sérgio Nunes†, Jorge Teixeira*† and Eugénio Oliveira*†
Faculdade de Engenharia da Universidade do Porto - DEI† - LIACC*
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
Emails: las@fe.up.pt, ssn@fe.up.pt, jft@fe.up.pt, eco@fe.up.pt

***Abstract*—We propose an unsupervised method for propagating automatically extracted fine-grained topic labels among news items to improve their topic description for subsequent text classification procedure. This method compares vector representations of news items and assigns to each news item the label of its closest neighbour with a different topic label. Results obtained show that high precision can be achieved in propagating the top ranked topic label, and that 2-gram and 3-gram feature representations optimize the precision.**

I. Introduction

News feed items are usually composed by a title, a short body (1-3 sentences) and some meta information, such as date, source or authorship. In some cases, these items also include a label that describes the *high-level topic* of the news – such as "Sports" or "Politics" – usually associated with a specific thematic section of the news source that published the feed. More focused topic information is *sometimes* found in the title of the news feed item, when the title begins by a label that provides additional fine-grained description of the topic (e.g. "*G20*: spotlight falls on police again"[1] or "*Centres de rétention*: la Cimade dénonce une "mascarade" de l'appel d'offre"[2]).

However, attempts to use such fine-grained topic labels for feed classification face several challenges. First, one can easily find hundreds or thousands of such very specific topic label in a given set of RSS (Really Simple Syndication) feeds, as opposed to only few dozen high-level topic tags that are usually associated with news sections. Second, many fine-grained topic tags can potentially be assigned to a given news source with different *scopes* of specialization. For example, for a news item describing a soccer match between two teams playing the European Champions League, it is possible to find in its title labels such as [Sports], [Soccer], [Champions League], or the name of any of the two teams. In other cases, title tags may express different, yet compatible, *perspectives* about the event covered in the news item. For example, a news source might label news regarding the pirate attacks on ships in the coast of Somalia either by the theme – [Piracy] or [Pirate attack] – by geography – [Somalia] or [Indian Ocean] – or by specific event – [Liberty Sun][3]. For classification algorithms this may represent a problem, since *very similar* items (i.e. RSS feeds from different news source reporting the same event) are considered to belong to different classes (i.e. were assigned different topic/class labels). This problem is worsened by the fact that when labeling with such fine-grained topic labels, it becomes very difficult to maintain a consistent topic assignment policy among annotators over time, even inside the *same* news source. If multiple spelling conventions or (domain dependent) synonyms are used, then the classification problem becomes even harder since multiple "equivalent" classes are labeled differently.

In this paper we present an approach that tries to reduce the impact of such topic label fragmentation for classification purposes. Our motivation comes from previous work done on quotation extraction from on-line news feeds [1]. In this context, extracted quotations are classified into specific topics (i.e. *someone* said *something* about a *specific topic*), and the fine-grained topic labels found in news titles are used to train the topic classifier. Because of the previously described problems, only the larger classes are kept (i.e. those associated with more frequent topic labels) during training and classification. However, due to the Zipfian-like nature of topic distribution, this strategy leaves out many interesting topics. In a nutshell, we want to investigate if we can automatically assign additional topic labels to a given news feed in order to improve the description of the feed with labels related with alternative but equally valid *scopes* or *perspectives*. We can later use such additional labels for training *multiple* classes thus increasing the number of *positive examples* for the classes found to be compatible. Additionally, by excluding the "equivalent" items from the set of *negative examples* for the newly found compatible classes, we will be able to reduce the number of "incorrect" negative examples used while learning the classification model of such classes. Hopefully, this will allow us to improve the precision of the quotation classification procedure, and avoid the need for truncating the number of topics used.

II. Related Work

Most of the work that has some connection to ours has been developed in the field of Topic Detection and Tracking. Pons-Porrata et al. [2] propose an incremental hierarchical clustering algorithm for news with the purpose of organizing news items both in a hierarchy that includes news *topics* (e.g.

[1]Taken from the http://www.guardian.co.uk.
[2]Taken from the http://www.lemonde.fr.
[3]Liberty Sun is a U.S.-flagged cargo ship attacked by Somali pirates in April 2009.

"Kosovo - Peace Agreement") and *events* that occur under such topics (e.g. "Agreement Sign"). News items are represented by a vector containing three types of features: (i) document terms, (ii) temporal references (dates) and (iii) geo-entities. Document similarity is computed by a function that explicitly takes into account document content and temporal-spatial proximity. Evaluation of the clustering algorithm is performed on a collection of 452 newspaper documents covering 48 topic and 68 non-unitary events. Results show that temporal information is beneficial for the news clustering process while spatial information tends to generate noisy results.

An algorithm for performing *unsupervised* discovery of topics labels from news is presented by Sista et al. [3]. The system is trained in three steps. First, it finds *descriptive phrases* in the collection of training documents, including names of entities (people, places and organizations) and n-grams with high level of lexical cohesion. Second, an initial set of topic labels is assigned to each training document based on the descriptive phrases they contain. Finally, the Estimate-Maximize procedure is used to find the *support words* associated with each topic. Final topic label assignment is performed using these support words. Thus, a topic label can be assigned to a document that does not include it. Evaluation was performed by manually computing label assignment precision at ranks 1-5 for a sample of 100 documents from two collections of newspaper documents in English and Arabic. Precision values ranged from 96% to 82% for English and 88% to 75% for Arabic. Entity name features were found to contribute only marginally to overall precision.

Bigi et al. [4] compare five statistic models for topic identification on newspaper text using a text-classification framework approach. Models were trained using a 80M words corpus divided in seven topics, and each news item had either one or *two* manually assigned tag. The best results in topic propagation were achieved with an unigram "cache" model that is dynamically updated over time. However, results for all five models in assigning the two correct tags to the corresponding texts were quite modest (precision $< 45\%$).

Overall, our work differs from existing research since we are explicitly developing an unsupervised pre-processing step for optimizing subsequent text classification procedures. As far as we know, this is a novel approach since most news classification systems avoid dealing with such a large number of classes by truncating them to the top frequent ones. Also, contrary to most related works, we are processing very short texts (40-60 words), while dealing with thousands of classes.

III. Propagating Topic Tags

We will try to explore the fact that news about the same topics should have similar contents. This should be specially so for news provided by different sources but covering the same event. If two news feeds items are found to have *very similar* content but have lexically different topic labels, we might assume that the topic label of each feed item can be propagated to the other, so that both of them will have one additional (hopefully valid) topic label.

More formally, let $\mathcal{F} = \{f_1, f_2, ...f_{|\mathcal{F}|}\}$ be a set of $|\mathcal{F}|$ news feed items. Each feed item, $f_i = [l_i, t_i, b_i]$, is composed of a *topic label*, l_i, a *news title*, t_i, and a *short text body* b_i (we will ignore related metadata in this work). Let $\overline{f_i}$ be a vector representation of the *content* of the feed item f_i, which includes features extracted from both the title t_i and the body b_i. For example, $\overline{f_i}$ can be a *bag-of-words* representation of t_i and b_i, with features weighted by tf-idf [5] over the feed set $\mathcal{F}$. The content of two news feeds, f_i and f_j, can be compared by using a vector similarity metric over the corresponding vector representations, $\overline{f_i}$ and $\overline{f_j}$. Using such vector representation and similarity metric, it is possible to obtain all the pairwise similarities between news feeds in $\mathcal{F}$.

Let $\mathcal{N}(f_i)$ be the ordered set of neighbours of f_i, i.e. the set of feeds item in $\mathcal{F}$ ordered by degree of similarity with f_i. Let $\mathcal{N}_1(f_i)$ be the nearest neighbour of f_i in such metric space. $\mathcal{N}_2(f_i)$ will be the second nearest neighbour, $\mathcal{N}_3(f_i)$ the third and so on. For each feed item f_i, let l_{i1}, l_{i2} ... l_{ik} be the topic labels associated with each of its neighbours, $\mathcal{N}_1(f_i)$, $\mathcal{N}_2(f_i)$ ... $\mathcal{N}_k(f_i)$. We will define l_i^{top} as the label of the closest neighbour of f_i whose label is *different* from l_i, given that the *inter-item distance* (or *inter-item similarity*) is lower (or higher) than a pre-defined threshold. Thus, l_i^{top} can be considered the best option for assigning a *new* topic label to f_i. Our method consists in propagating the l_i^{top} to f_i. We will not immediately propagate l_i back to the closest neighbour of f_i, f_i^{ngbr}, because f_i might not be its closest neighbour (there might be other feed items closer to f_i^{ngbr}).

We could increase the number of new topic labels assigned to each feed item by propagating more labels than just the first different label found in the neighbours list (e.g. the top 3). Alternatively, we could propagate only the top label but repeat the propagation process iteratively, so that a new topic label that was assigned to a given feed item in one iteration, could be propagated to one of its neighbours in the next iteration. However, in this work we wish to evaluate the usefulness of the basic propagation principle itself so we will only propagate one label (l_i^{top}) executing only one iteration. In other words, we will only perform and evaluate the *new nearest-neighbour topic label*.

IV. Experimental Set-Up

We collected RSS feeds from eight distinct mainstream portuguese news sources for a period of about 6 months (mid November 2008 to mid April 2009). This allowed us to collect 90,780 feeds items. Among these, a significant fraction (30.4%) had titles with the structure we wish to explore, i.e. "[topic label]: remainder of the title...". The use of this pattern in titles varied greatly, ranging from less than 1% in two cases to more than 70% in one news source. We performed additional filtering using a dictionary containing names of entities that are frequently mentioned in news to exclude cases where the title actually refers to a quotation (e.g. "Obama: Economy improving, crisis not over"). In the end we obtained 18,309 news feeds containing a valid topic label corresponding to 3,082 different topic labels ranging from

very frequent and generic labels, such as "Futebol/Soccer" (886 items), "Música/Music" (393) or "EUA/USA" (386)", to long tail labels that were found in only one document such as "Cogumelos silvestres/Wild Mushrooms", "Grammy Awards" or "Botox".

News feed items were vectorized by generating n-gram features from the title and the body (we differentiate features coming from the title from n-grams coming from the body). In our experiments we explored 4 different types of features: (i) unigram features, (ii) 2-gram features, (iii) 3-gram features and (iv) 4-gram features. With these four options we wish to measure the balance between a more compact feature set which ignores all word ordering information (i.e. unigram features) and order sensitive yet much sparser feature set (i.e. 4-gram features). In any case, these four options represent a rather straight-forward approach, which does not require any more sophisticated language pre-processing. Features were weighted by tf-idf to demote those that occur in many news items. The resulting vectors were compared using the cosine metric [5].

We evaluate the propagation of topic label by manually comparing l_i^{top} with the originally assigned label l_i and consider 5 different possibly correct (C) cases:

- C1: Different Perspective – l_i^{top} addresses an alternative perspective of the news (e.g. location vs. time);
- C2: Generalization/Specialization – l_i^{top} is a generalization or a specialization of the concept described by l_i (e.g. "Sports" vs. "Soccer");
- C3: Non-Obvious Synonym – l_i^{top} is equivalent to l_i but there is no lexical intersection (i.e. words in common) between both labels that would suggest such equivalence beforehand.
- C4: Obvious Synonym – l_i^{top} is equivalent to l_i, but there is sufficient lexical overlap between both to make such equivalence rather obvious (e.g. l_i^{top} is lexically included in l_i).
- C5: Spelling Variation – l_i^{top} and l_i differ only in minor spelling variations (sometimes one is a misspelled version of the other).

Deciding cases C1, C2 and C3 may require consulting external sources (for example the corresponding complete news item). All other possibilities not listed before are considered incorrect (I).

We compared the results of the label propagation algorithm using the four different options for generating vector features (unigrams, 2-grams, 3-grams and 4-grams). The four runs were configured to obtain almost equal recall figures so that the corresponding values of the precision regarding l_i^{top} could be compared. The relation between desired recall and precision can be controlled by setting a threshold on the minimum value of similarity between feature vectors. The larger this threshold, the less probable is to obtain false positives – which can result from noisy or ambiguous features – and the higher should be the value of precision at the cost of recall. We are mostly interested in looking at the high-precision section of the precision-recall curve since l_i^{top} is intended to be used as class label in subsequent text classification procedures.

V. Results and Analysis

Table I presents results (in %) regarding the propagation of l_i^{top} (cases C1 to C5 described before), for each of the four feature generation options experimented. Manual evaluation was performed over a random sample of 30% of the resulting topic label attributions. The position on the recall vs. precision curve was controlled by manually setting the minimum inter-item similarity threshold parameter in order to achieve high precision ($\simeq$90%) at comparable recall values. The minimum value for the inter-item similarity (i.e. cosine) varied from to 0.2 (for unigram features) to 0.33 (when using 4-gram features). The resulting recall values obtained oscillated between 7.1% and 7.2% for all the runs (approximately to 1,310 l_i^{top} label assignments in a set of 18,309 news feeds), which allowed a fair comparison of the corresponding precision values achieved.

TABLE I
Precision (in %) at Recall 7.1%-7.2%.

	unigram	2-gram	3-gram	4-gram
C1	24.9	31.1	33.9	30.7
C2	42.0	39.6	35.5	30.7
C3	6.3	4.8	4.6	6.1
C4	8.7	8.0	11.6	11.6
C5	7.9	10.4	8.3	9.7
$\sum$C	89.8	93.9	93.8	88.7
I	10.2	6.1	6.2	11.3

Results show that for equivalent recall figures, precision achieved by using 2-gram and 3-gram features is significantly higher than that obtained using either unigram or 4-gram features. This result is important since it is very common to use unigram features in several news classification procedures, when, as seen, 2-gram and 3-gram features seem to carry more information about the topics. There may be several reasons for this, but the fact that 2-gram and 3-gram features keep the names of most entities mentioned in news feeds intact (specially names of people and organizations) may be particularly significant. On the other hand, on such small text snippets, 4-gram features tend to generate a too sparse feature space which affects the efficiency of feature vector comparison, since it tends to relatively promote the importance of topic-unrelated features, such as common stylistic formulations or language specific fixed-expressions. Also, most label propagations added information either about different *perspectives* (case C1) or different *scopes*. Still for about 5% of the cases the methods propagated *non-obvious synonyms* (C3).

Further analysis of incorrect label propagation revealed two major type of errors, which were common to the four runs. The first, and by far the most common, was related to news about local events. These news usually refer to accidents (e.g. car accident or a burning house) or crimes events (e.g. a shop robbery) that occurred in a specific city or neighbourhood

and are usually labeled using the name of that location. In many cases, the topic label that was propagated to these feeds items is actually the name of *another* and unrelated location (e.g. another city). The reason for this lies in the fact that, apart from mentioning different locations, most news about accidents and crime events at local level are actually quite similar in structure and content, i.e. they report very similar occurrences using similar wordings. Also, they include mentions to certain typical entities (e.g. the Police or the Fire Department) and describe common standard procedures (e.g. an arrest). This specific type of errors can probably be avoided by using a geographical ontology to identify such news and either excluding the assignment of incompatible labels (e.g. by taking into account distance of the locations) or by only propagating labels that refer to different perspectives (e.g. "Crime").

The second most frequent type of error was the propagation of an incorrect yet related label. For example, news about the *European* Euromillion Loto received labels that are associated with the *Portuguese* Loto (which is independent of the european one), or news about a specific soccer club or player were assigned a label about another soccer club or player. This type of errors seems to be more difficult to solve and might require looking to labels other than the top suggested one, l_i^{top}, and using more complex criteria to select the correct label (e.g. minimizing the distance to *several* news feed items *simultaneously*).

VI. Conclusion and Future Work

In this paper we showed that it is in fact possible to automatically propagate topic label between news items using an unsupervised process based on the content of the news. The majority of labels propagated add relevant *perspective* or *scope* information to items. We also showed that 2-gram and 3-gram features seem to carry more topic-related information, and can thus be used in subsequent classification tasks.

Future work will address limitations concerning recall. One option is to change the label ranking procedure so that information coming from multiple neighbours is combined: if a new label is found in *several* of the nearest neighbours of an item, then it might be considered a good topic suggestion even if it is not the top ranked one. We can also perform multiple iterations of the algorithm to propagate novel label information to nodes which could already be close enough to their neighbours but had the *same* label.

Also, since we are already capable of providing multiple tags to a single news item, we will try to mine *association rules* between topic labels using strategies similar to those described by Heymann et al. [6]. This will allow us to obtain *networks* of topic labels that can be used as background knowledge in future version of our method, or used as gold-standard in automatic evaluation procedures. Finally, for allowing our method to scale to larger datasets, we wish to experiment high-performance methods for finding the top-k nearest-neighbours, such as the Min-Hash algorithm [7].

Acknowledgments

Work partially supported by grants SFRH/BD/23590/2005 and SFRH/BD/31043/2006 from FCT Portugal, and by grant SAPO/BI/UP/2009 from Portugal Telecom.

References

[1] L. Sarmento, S. Nunes, and E. Oliveira, "Automatic extraction of quotes and topics from news feeds," in *4th Doctoral Symposium on Informatics Engineering*, Porto, Portugal, 2009.

[2] A. Pons-Porrata, R. Berlanga-Llavori, and J. Ruiz-Shulcloper, "Building a hierarchy of events and topics for newspaper digital libraries," in *Proceedings of* 25^{th} *European Conference on IR Research (ECIR'03)*, F. Sebastiani, Ed., vol. 2633. Pisa, Italy: Springer-Verlag, 2003, pp. 588–596.

[3] S. Sista, R. Schwartz, T. R. Leek, and J. Makhoul, "An algorithm for unsupervised topic discovery from broadcast news stories," in *Proceedings of the 2nd international Conference on Human Language Technology Research*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 110–114.

[4] B. Big, A. Brun, I. Zitouni, J. Haton, and K. Smaïli, "A comparative study of topic identification on newspaper and e-mail," *String Processing and Information Retrieval, International Symposium on*, vol. 0, p. 0238, 2001.

[5] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[6] P. Heymann, D. Ramage, and H. Garcia-Molina, "Social tag prediction," in *31st Annual International ACM SIGIR Conference (SIGIR'08)*, July 2008.

[7] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *The VLDB Journal*, 1999, pp. 518–529. [Online]. Available: http://citeseer.ist.psu.edu/gionis97similarity.html